# DES ARCHIVES DU WEB AUX DONNÉES
## Une décennie de nouveaux services et collaborations

*Valérie Schafer*

Professeur d'histoire européenne contemporaine, université du Luxembourg
valerie.schafer@uni.lu

*Sophie Gebeil*

Maître de conférences en histoire contemporaine, université Aix-Marseille
sophie.gebeil@univ-amu.fr

La mise à disposition pour la recherche de données issues du Web archivé correspond à une tendance engagée au sein des institutions patrimoniales depuis les années 2010. Elle se matérialise à l'échelle européenne puis internationale de manière assez inédite à la faveur d'une étude sur l'archivage des traces numériques de la crise de la COVID-19 à partir de 2020. En explorant l'émergence de nouveaux services et collaborations lors de projets mettant en jeu la fourniture des données issues du Web archivé sur la dernière décennie, en particulier à l'Institut national de l'audiovisuel (INA) et à la Bibliothèque nationale de France (BnF), cet article invite à penser les transformations des relations entre chercheurs et institutions patrimoniales. Après avoir souligné à quel point cette évolution est le résultat d'efforts engagés depuis plusieurs années en termes d'indexation, d'accès, de fourniture d'outils, nous analysons les modalités et enjeux de la co-construction d'une recherche sur les données du Web archivé, notamment pour la constitution et documentation des corpus et leur lecture distante.

*Mots-clefs : Web archivé, données, corpus, institutions patrimoniales, lecture distante*

*Data provision from the archived Web is a trend that has been underway within heritage institutions since the 2010s. It has been materialized on a European and then international scale in a rather unprecedented way in a study on the digital traces of the COVID-19 crisis from 2020. By exploring the emergence of new services and collaborations in projects involving the release of data from the archived Web over the last decade, notably at the Institut national de l'audiovisuel and at the Bibliothèque nationale de France, this article aims to analyse the transformations in the relationship between researchers and heritage institutions. After underlining the fact that this evolution is the result of efforts*

*undertaken for several years in terms of indexing, accessibility, and provision of tools, we explore the challenges and the co-construction that research projects on data from archived Web involve, may it be for the creation and documentation of corpora and their distant reading.*

*Keywords: Archived Web, data, corpora, heritage institutions, distant reading*

Alors que la crise de la COVID-19 s'étend en Europe à partir de mars 2020, la nature exceptionnelle et globale de la pandémie incite les institutions qui archivent le Web à créer des collections spéciales dédiées à cet événement, dont on ignore encore la portée, la durée et les effets. Déjà, des collectes de contenus du Web couvrant des événements exceptionnels, prévus (Jeux olympiques, élections, etc.) ou imprévus (attaques terroristes de 2015-2016, mouvement Nuit debout ou MeToo), avaient été entreprises par ces institutions en complément de leurs collectes larges, notamment en France. Toutefois, avec la crise de la COVID-19 un mouvement simultané d'archivage du Web d'ampleur internationale est entrepris, qui suscite aussi l'intérêt de l'International Internet Preservation Consortium (IIPC)[1]. L'IIPC invite notamment ses membres, soit essentiellement des bibliothèques nationales et des institutions qui archivent le Web, à remonter des sélections de sites et d'URLs pour constituer une vaste collection[2]. Ce mouvement qui débute au premier semestre 2020 est quasi simultané du lancement d'un projet collaboratif entre des archivistes européens du Web et des chercheurs intéressés aux archives du Web, le projet WARCnet (Brügger, 2020)[3]. Il va prendre en partie pour objet d'étude cet effort collectif et permet, à la faveur de collaborations étroites avec le monde des archives, d'accéder aux données, métadonnées (Pomerantz, 2015) et données dérivées de ces collectes spéciales. Dans bien des institutions, les contenus des archives du Web ne sont accessibles que dans leurs espaces de consultation (dans le cas français, la Bibliothèque nationale de France [BnF] et l'Institut national de l'audiovisuel [INA]), notamment pour des raisons de droit d'auteur. Obtenir la fourniture de métadonnées et de données dérivées en masse (liste des URLs archivées, date de la collecte, heure, récurrence, nature des sites, etc.) liées aux archives du Web de la COVID-19[4]

1. Sur l'IIPC, voir https://netpreserve.org.
2. https://archive-it.org/collections/13529.
3. https://cc.au.dk/en/warcnet.
4. L'équipe WARCnet a notamment reçu des données de la BnF et de l'INA, mais aussi de la British Library, de la Bibliothèque royale du Danemark ou encore de la Bibliothèque nationale du Luxembourg, pour n'en citer que quelques-unes.

permet d'envisager un premier traitement transnational via des lectures distantes. Ce mouvement vers une recherche fondée sur les masses de données du Web archivé s'amplifie avec l'étude plus précise de la vaste collection de la COVID-19 rassemblée par l'IICP (Bingham, 2020), en collaboration avec l'équipe canadienne du projet Archives Unleashed (Ruest *et al.*, 2021), ou encore un projet d'un an Buzz-F avec le BnF DataLab portant sur la viralité en ligne dans une perspective diachronique[5]. Ces projets mais aussi la création de « labs » (le BnF DataLab est inauguré en 2021 – voir Carlin et Laborderie, 2021 ; l'INA a lancé le sien à la fin 2022[6]) invitent à penser, à la faveur de l'émergence d'un « Web des données », les transformations des relations entre institutions patrimoniales et chercheurs, et le nécessaire accompagnement de ces derniers, qu'il concerne le rapport aux sources et la création de corpus ou leur outillage, intellectuel et computationnel.

Ces nouveaux services, dispositifs d'accompagnement et fonctions des « lieux de savoir » s'inscrivent dans une tendance plus longue de fourniture de données liées à leurs collections, engagée pour le Web archivé dès les années 2010. En faisant évoluer les logiques de collecte, de collections et d'indexation, en plaçant de plus en plus au cœur de leur approche les données, en des logiques documentaires différentes, les institutions influencent les recherches qui peuvent être menées à partir des archives du Web. Après un retour sur cette tendance engagée au cours de la décennie 2010, qui considère de plus en plus les données du Web archivé, nous insisterons sur le tournant de 2020, la coconstruction et les enjeux qu'impliquent ces nouvelles approches.

## QUAND LE WEB ARCHIVÉ SE FAIT DONNÉES… DES ÉTAPES IMPORTANTES DÈS LA DÉCENNIE 2010

En France comme à l'international, les chercheurs qui s'intéressent aux archives du Web mettent d'abord l'accent sur des défis méthodologiques liés à un patrimoine nativement numérique en cours de constitution rapide et cherchent à affiner les approches de ce que Niels Brügger qualifie de « reborn digital heritage » (2018). C'est davantage du côté de la lecture qualitative que s'orientent les chercheurs européens. Il y a plusieurs raisons à cela, à commencer par l'impossibilité d'accéder aisément à des données autres que les pages et contenus web archivés, que ce soit par l'interface de la Wayback Machine[7] d'Internet Archive ou dans les institutions nationales. Il y a aussi le souhait de

5. https://www.bnf.fr/fr/les-projets-de-recherche.
6. https://www.ina.fr/actualites-ina/l-ina-lance-le-lab-la-data-media-au-service-des-chercheurs.
7. https://archive.org/web.

prendre la mesure des spécificités de ces sources pour lesquelles les méthodes numériques issues du Web vivant peinent à s'appliquer. Toutefois, quelques projets ont déjà une approche des traces numériques par la lecture distante, c'est-à-dire fondée sur des analyses de données à l'aide d'outils informatiques (Moretti, 2007). C'est le cas dans le projet e-diasporas soutenu par l'Agence nationale de la recherche (ANR) de 2008 à 2012 de Dana Diminescu[8], ou dans le travail sur les mémoires de la Grande Guerre telles qu'elles peuvent être saisies par les archives du Web (Clavert, 2018). Sur ce même thème, Valérie Beaudouin cherche par exemple à tirer parti des hyperliens pour l'analyse de réseaux (Beaudouin, 2019) et travaille sur des métadonnées en étant accompagnée dans ces démarches par la BnF (Beaudouin, 2018). Cette dernière réalise également dans le cadre d'une convention des collectes spécifiques pour le projet à partir de 2013 (Sandras et Stirling, in Beaudouin *et al.*, 2018, p. 91-105).

## L'enjeu des collectes, de l'accès et de la recherchabilité

L'une des préoccupations de la décennie 2010 est en effet fortement liée à la constitution des corpus et à la consultation qui se fait majoritairement dans les institutions, pour le chercheur qui ne se cantonne pas à une recherche en ligne via la Wayback Machine d'Internet Archive.

Comme dans le cas de l'étude de la Grande Guerre sur le Web, le travail de Sophie Gebeil[9] sur les mémoires de l'immigration en ligne implique une étroite collaboration avec les archivistes pour retrouver d'anciens sites ou préserver des pages sélectionnées en ligne : la BnF et l'INA ont des périmètres d'archivage qui ne prennent pas en compte avec la même récurrence ou profondeur tous les contenus et certains sites peuvent échapper à la collecte ou être collectés très sporadiquement. Aussi, avec un statut de chercheur associé à la BnF, elle peut au cours de la recherche compléter les collections via l'interface BCWeb et proposer des sites en vue de collectes ciblées, et ce, en étroite relation avec le Dépôt légal numérique (DLN) et le Département Philosophie, histoire et sciences humaines (fig. 1). Ainsi s'amorcent une collaboration et un enrichissement des collections à des fins de recherche spécifiques à la demande du chercheur. Précédemment, des collectes électorales avaient, notamment, associé des équipes de recherche aux définitions des paramètres de collecte.

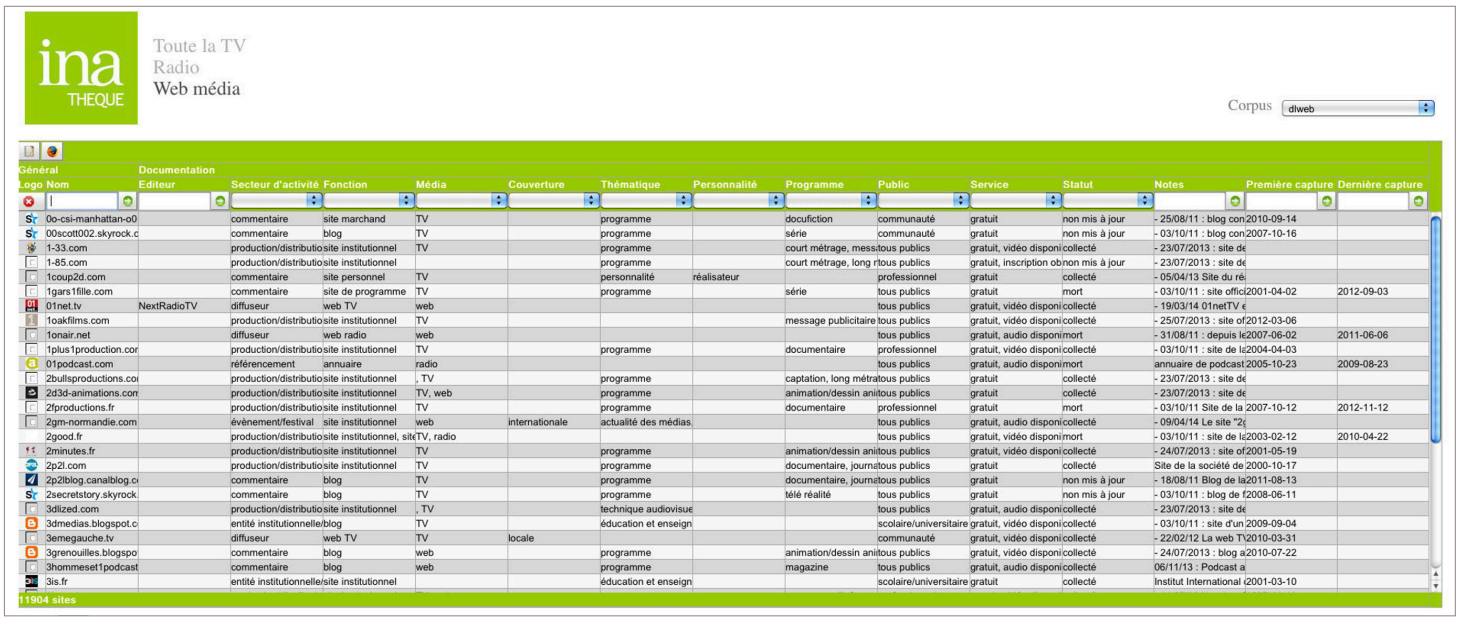
8. http://www.e-diasporas.fr.
9. Dans le cadre de sa thèse entamée en septembre 2010.

Figure 1. Capture d'écran de l'application BnF Collecte du Web (BCWeb), 29 mars 2015


*Légende : BCWeb permet à l'utilisateur qui dispose d'un compte personnalisé, de suggérer des URLs en vue des collectes ciblées. Source : © BnF.*

Identifier des pages ou des contenus web déjà archivés qui soient pertinents est encore peu aisé car les interfaces ne permettent alors ni la recherche plein texte ni par mot-clé. Il faut utiliser des tableurs comprenant les listes des sites collectés par les institutions, directement intégrés dans le navigateur de l'INA (fig. 2) ou partagés par l'équipe du DLN dans le cadre du contrat de chercheur associé avec la BnF. Ce sont des outils parfois pensés d'abord pour des usages internes aux institutions qui sont mis à disposition. Ces listes permettent toutefois de prendre la mesure des contenus conservés, alors que la Wayback Machine d'Internet Archive ne permet qu'une recherche par URL, sans vision globale des contenus existants. L'identification même des sources constitue ainsi un défi méthodologique à part entière, avant que ne s'améliorent les conditions de recherchabilité dans le Web archivé.

## Figure 2. Capture d'écran du navigateur de consultation réalisée le 15 mai 2014




| Général | | Documentation | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Logo | Nom | Editeur | Secteur d'activité | Fonction | Média | Couverture | Thématique | Personnalité | Programme | Public | Service | Statut | Notes | Première capture | Dernière capture | |
| | 0o-csi-manhattan-o0 | | commentaire | site marchand | TV | | programme | | docufiction | communauté | gratuit | non mis à jour | - 25/08/11 : blog con | 2010-09-14 | | |
| | 00scott002.skyrock.c | | commentaire | blog | TV | | programme | | série | communauté | gratuit | non mis à jour | - 03/10/11 : blog con | 2007-10-16 | | |
| | 1-33.com | | production/distributio | site institutionnel | TV | | programme | | court métrage, messa | tous publics | gratuit, vidéo disponi | collecté | - 23/07/2013 : site de | | | |
| | 1-85.com | | production/distributio | site institutionnel | | | programme | | court métrage, long r | tous publics | gratuit, inscription ob | non mis à jour | - 23/07/2013 : site de | | | |
| | 1coup2d.com | | commentaire | site personnel | TV | | personnalité | réalisateur | | professionnel | gratuit | collecté | - 05/04/13 Site du ré | | | |
| | 1gars1fille.com | | commentaire | site de programme | TV | | programme | | série | tous publics | gratuit | mort | - 03/10/11 : site offici | 2001-04-02 | 2012-09-03 | |
| | 01net.tv | NextRadioTV | diffuseur | web TV | web | | | | | tous publics | gratuit, vidéo disponi | collecté | - 19/03/14 01netTV é | | | |
| | 1oakfilms.com | | production/distributio | site institutionnel | TV | | | | message publicitaire | tous publics | gratuit, vidéo disponi | collecté | - 25/07/2013 : site of | 2012-03-06 | | |
| | 1onair.net | | diffuseur | web radio | web | | | | | tous publics | gratuit, audio disponi | mort | - 31/08/11 : depuis le | 2007-06-02 | 2011-06-06 | |
| | 1plus1production.cor | | production/distributio | site institutionnel | TV | | programme | | documentaire | professionnel | gratuit, vidéo disponi | collecté | - 03/10/11 : site de la | 2004-04-03 | | |
| | 01podcast.com | | référencement | annuaire | radio | | | | | tous publics | gratuit, audio disponi | mort | annuaire de podcast | 2005-10-23 | 2009-08-23 | |
| | 2bullsproductions.co | | production/distributio | site institutionnel | , TV | | programme | | captation, long métra | tous publics | gratuit | collecté | - 23/07/2013 : site de | | | |
| | 2d3d-animations.com | | production/distributio | site institutionnel | TV, web | | programme | | animation/dessin ani | tous publics | gratuit | collecté | - 23/07/2013 : site de | | | |
| | 2fproductions.fr | | production/distributio | site institutionnel | TV | | programme | | documentaire | professionnel | gratuit | mort | - 03/10/11 Site de la | 2007-10-12 | 2012-11-12 | |
| | 2gm-normandie.com | | événement/festival | site institutionnel | web | internationale | actualité des médias | | | tous publics | gratuit, audio disponi | collecté | - 09/04/14 Le site "2 | | | |
| | 2good.fr | | production/distributio | site institutionnel, site | TV, radio | | | | | tous publics | gratuit, vidéo disponi | mort | - 03/10/11 : site de la | 2003-02-12 | 2010-04-22 | |
| | 2minutes.fr | | production/distributio | site institutionnel | TV | | programme | | animation/dessin ani | tous publics | gratuit, vidéo disponi | collecté | - 24/07/2013 : site of | 2001-05-19 | | |
| | 2p2l.com | | production/distributio | site institutionnel | TV | | programme | | documentaire, journa | tous publics | gratuit | collecté | Site de la société de | 2000-10-17 | | |
| | 2p2lblog.canalblog.c | | commentaire | blog | TV | | programme | | documentaire, journa | tous publics | gratuit | non mis à jour | - 18/08/11 Blog de la | 2011-08-13 | | |
| | 2secretstory.skyrock. | | commentaire | blog | TV | | programme | | télé réalité | tous publics | gratuit | non mis à jour | - 03/10/11 : blog de f | 2008-06-11 | | |
| | 3dlized.com | | production/distributio | site institutionnel | , TV | | technique audiovisue | | | tous publics | gratuit, audio disponi | collecté | - 23/07/2013 : site de | | | |
| | 3dmedias.blogspot.c | | entité institutionnelle | blog | TV | | éducation et enseign | | | scolaire/universitaire | gratuit, vidéo disponi | collecté | - 03/10/11 : site d'un | 2009-09-04 | | |
| | 3emegauche.tv | | diffuseur | web TV | TV | locale | | | | communauté | gratuit, vidéo disponi | collecté | - 22/02/12 La web T\ | 2010-03-31 | | |
| | 3grenouilles.blogspo | | commentaire | blog | web | | programme | | animation/dessin ani | tous publics | gratuit, vidéo disponi | collecté | - 24/07/2013 : blog a | 2010-07-22 | | |
| | 3hommeset1podcast | | commentaire | blog | web | | programme | | magazine | tous publics | gratuit, audio disponi | collecté | 06/11/13 : Podcast a | | | |
| | 3is.fr | | entité institutionnelle | site institutionnel | | | éducation et enseign | | | scolaire/universitaire | gratuit | collecté | Institut International | 2001-03-10 | | |

11904 sites

*Légende : Accessible via le navigateur de consultation de l'INA depuis 2012, le tableur permet à l'usager de chercher des contenus collectés au sein de la base de données selon différents critères (URL, secteur, type de média, public cible, etc.). Source : © INA.*

Les chercheurs intéressés aux archives du Web étant peu nombreux, ils ont toutefois la possibilité de nouer un dialogue privilégié avec les institutions d'archivage qui, dans le même temps, entament une réflexion sur les données, comme le relève Emmanuelle Bermès pour le cas de la BnF :

> *Le point de contact entre bibliothèque et chercheurs, entre collections et corpus, se situe dans la coconstruction de nouveaux ensembles de données, comme dans le cas du projet « Le devenir du patrimoine numérisé en ligne : l'exemple de la Grande Guerre », où une collecte d'archives du Web fut réalisée spécifiquement pour permettre l'analyse des réseaux amateurs et institutionnels sur le Web. Lorsqu'en 2014, le Labex Obvil émet la demande d'obtenir une copie de Gallica – alors riche de 3 millions de documents – pour effectuer des recherches relevant du « big data », il devient clair que la bibliothèque doit s'organiser pour répondre de manière fluide à ce type de demande. C'est ainsi qu'est imaginé le projet DSR[10]-Corpus, inscrit au plan quadriennal de la recherche de l'établissement pour la période 2016-2019, qui vise à préfigurer de nouveaux services pour les chercheurs autour des collections numériques. […] Dans le cadre du projet DSR-Corpus, à travers le site API[11] et données ou lors des hackathons, les collections numériques de la BnF sont exploitées suivant un nouveau paradigme, caractérisé dans le monde anglo-saxon par le terme « collections as data » […]. (Bermès, 2020, p. 57 et ss.).*

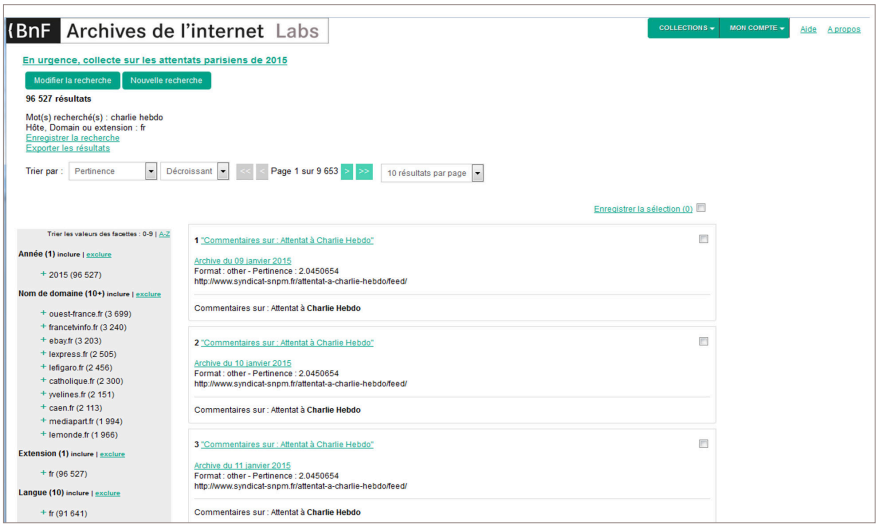
Au début du projet ANR Web90[12], dédié au Web des années 1990 et qui démarre en 2014, la recherchabilité des contenus est également un enjeu majeur. Il ne s'agit pas de créer un nouvel ensemble de données et de sites mais de travailler sur ceux préservés. Toutefois, à mi-parcours, le projet Web90 va bénéficier du projet DSR-Corpus et de l'implémentation du plein texte dans une partie des archives du Web de la BnF et notamment dans celles des années 1990. L'indexation est importante pour le projet mais aussi plus généralement pour la fourniture de services par la bibliothèque : ainsi pour les années 1990, la BnF conserve des collections qui ont été récupérées d'Internet Archive, avant qu'Internet et le Web n'entrent dans le dépôt légal de la BnF en 2006. Dès lors, la plus-value que peut offrir la BnF réside dans l'accès et la recherchabilité des contenus, non dans les contenus eux-mêmes, que l'on peut trouver en ligne par la Wayback Machine, mais qui sont alors recherchables uniquement par URLs. C'est une étape pour le passage du « patri-

10. Direction des services et des réseaux.
11. Application programming interface ou interface de programmation d'application.
12. https://web90.hypotheses.org/le-projet-web90.

moine comme stock » au « patrimoine valorisé » (Treleani, 2017). Un certain nombre d'éléments statistiques sur les collections indexées est également fourni, de manière encore assez globale (nombre de fichiers WARC[13], d'URLs collectées), en une mise à disposition de données déjà utilisées au sein de l'institution, mais qu'elle ne partageait pas forcément avec les chercheurs (qui ne les réclamaient pas non plus). Les données sont là, mais leur valeur était davantage exploitée en interne.

En parallèle, l'équipe Web90 peut suivre et réagir aux réflexions des professionnels sur la création de nouvelles interfaces de requête, la mise en forme des résultats, les facettes permettant leur tri (fig. 3).

Figure 3. Capture d'écran de l'interface de la BnF, 2016




*Légende : L'interface de la BnF permet une exploration plein texte et par facettes dans les archives du Web en 2016. Source : © BnF.*

Les institutions doivent composer avec la nouveauté de ces archives mais aussi leurs « traditions » internes, négocier entre logiques documentaires à privilégier pour le Web archivé, mais aussi l'ancrage dans des outils et approches existant au sein de l'institution. La BnF s'appuie sur l'interface Shine développée au sein des UK Web Archives pour développer sa propre interface de consultation du Web archivé, tout en cherchant à adapter celle-ci à ses visions et ses connaissances des expériences, pratiques et besoins de

13. Format Web ARChive, extension du format ARC approuvée en tant que norme internationale par l'International Organization for Standardization (ISO).

ses publics comme des services internes. Ainsi, des discussions portent par exemple sur la création d'un champ auteur dans l'interface de consultation. Même si la notion d'auteur est difficile à saisir pour les sites web, ce champ apparaît dans toutes les autres interfaces de requête de la BnF. Ces observations rendent palpables la gouvernance, les négociations (Schafer, Musiani, Borelli, 2016) et les contraintes institutionnelles. Elles permettent d'éviter un « déni d'interface » que relève Mélanie Roustan (2016), phénomène de naturalisation qui rend la structure invisible, mais aussi le travail qui se déploie en coulisse dans les institutions pour développer un accès et une offre de services.

## Des espaces de dialogue entre professionnels de l'archivage et chercheurs

En parallèle, d'autres espaces de dialogue entre chercheurs et professionnels de l'archivage et de la préservation existent déjà, en particulier au sein des ateliers du dépôt légal du Web de l'INA sous l'égide de Louise Merzeau et Claude Mussou (2017) : ils explorent une vaste palette de questions liées au patrimoine nativement numérique dès le début de la décennie 2010 (logiques d'usage et de redocumentarisation, migration de l'audiovisuel vers le Web, etc.). Les institutions mènent aussi des enquêtes, à l'instar de celle réalisée à la BnF (Moiraghi, 2018) dans le cadre du projet CORPUS, inscrit au plan quadriennal de recherche de l'établissement de 2016 à 2019. Les demandes des chercheurs en faveur d'un export des données, de captures d'écran des contenus du Web archivé, de démarches plus ouvertes en termes de partage sont récurrentes, au-delà de l'appréciation des services mis en place. Enfin, le dialogue commence également à l'échelle européenne, grâce à l'initiative de Niels Brügger de créer en 2012 le réseau européen RESAW[14] destiné à développer une infrastructure européenne d'accès aux archives du Web, ce qui n'est possible que dans un dialogue constant avec les institutions de préservation. Dans ce projet (qui ne se concrétise pas alors) sont déjà imaginés plusieurs niveaux d'interopérabilité et de partage des données, mais beaucoup d'institutions se heurtent à la question des droits d'auteur et cherchent davantage à fournir des outils internes pour accueillir les chercheurs qui viennent consulter les fonds dans les espaces institutionnels.

14. http://resaw.eu.

## Des outils de lecture distante

Si déjà des approches de lecture distante pionnières sont menées sur les archives du Web dans le cadre de projets collectifs (projet e-diaspora ou Grande Guerre), des projets individuels bénéficient aussi de cette approche et du développement d'outils d'analyse par les institutions. C'est le cas de la recherche sur les mémoires maghrébines de l'immigration précédemment évoquée. Ainsi en 2013 est menée une expérimentation de fouille avec l'équipe du DL Web de l'INA, à partir d'un outil d'extraction, « Proprioception »[15], développé par un de ses ingénieurs, David Rapin. Cette initiative est conduite à partir d'un nombre réduit de sites web traitant des mémoires de l'immigration maghrébine. Elle donne lieu à l'extraction, au format CSV[16], de centaines de milliers de liens sortants, du nombre d'images, de vidéos par page, ou de mots présélectionnés à partir d'une liste en vue d'une analyse lexicale (fig. 4a et 4b). Ces données permettent de cerner les évolutions des sites entre 2011 et 2012, dans la perspective d'une lecture encore largement exploratoire.

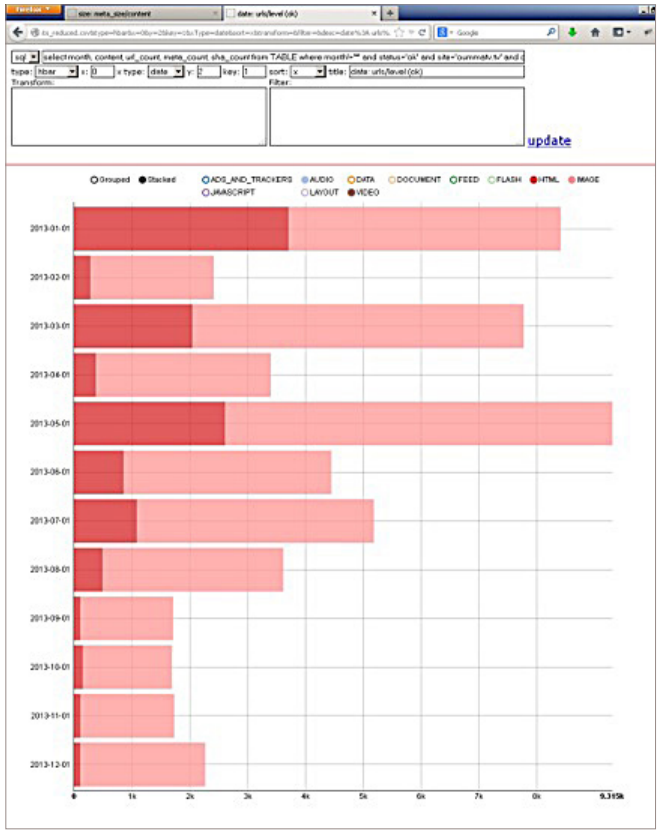
Figure 4a. Évolutions des liens sortants depuis le site Raspouteam.org entre décembre 2011 et mars 2012

| URL | Date | URL sortante | Nombre |
|---|---|---|---|
| raspouteam.org | 2011-12 | maps.google.com | 11 |
| raspouteam.org | 2011-12 | maps.google.fr | 26 |
| raspouteam.org | 2011-12 | twitter.com | 79 |
| raspouteam.org | 2011-12 | www.facebook.com | 80 |
| raspouteam.org | 2011-12 | www.raspouteam.org | 80 |
| raspouteam.org | 2012-01 | maps.google.com | 11 |
| raspouteam.org | 2012-01 | raspouteam.org | 21 |
| raspouteam.org | 2012-01 | maps.google.fr | 26 |
| raspouteam.org | 2012-01 | twitter.com | 109 |
| raspouteam.org | 2012-01 | www.facebook.com | 110 |
| raspouteam.org | 2012-01 | www.raspouteam.org | 111 |
| raspouteam.org | 2012-02 | maps.google.com | 11 |

*Légende : Extrait d'un premier traitement réalisé par David Rapin à partir de l'extraction des données issues des différentes versions du site Raspouteam.org archivées par l'INA. Source : © INA.*

15. Ce terme renvoie à l'idée de percevoir sans voir, ici transposée à l'exploration quantitative automatisée du corpus sans procéder à une lecture proche.
16. Comma-separated values ou valeurs séparées par des virgules.

Figure 4b. Évolution du nombre de pages HTML et d'images (fichiers .jpeg, jpg, .png, .tif, etc.) par mois au sein des différentes versions du site Oumma.tv archivé par l'INA

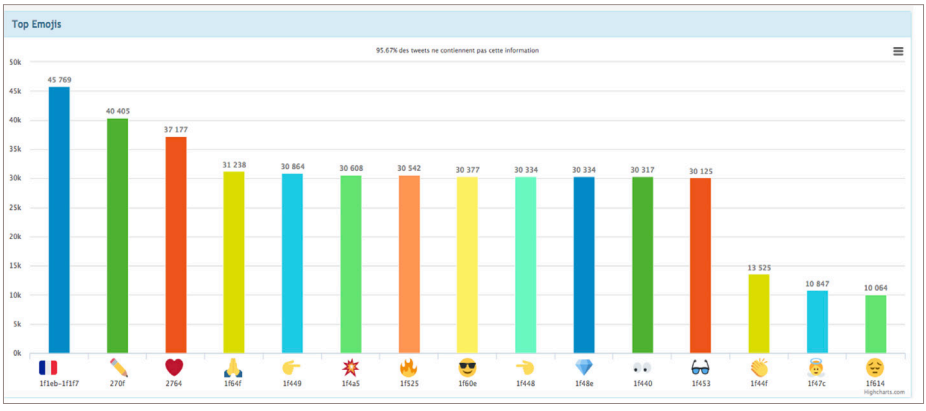



*Légende : Visualisation réalisée à partir des données issues des archives web de l'INA en 2013 en vue d'une première lecture distante de l'histoire du site. Source : © INA.*

Du côté de l'INA toujours, le périmètre de collecte lié à l'audiovisuel est élargi vers des plateformes en ligne, notamment celles de vidéo comme YouTube et Dailymotion. En 2014, c'est Twitter qui suscite à son tour l'intérêt (notamment pour suivre des comptes de journalistes ou de chaînes) et des interfaces spécifiques sont créées pour l'analyse des tweets que l'institut commence à collecter avec régularité. En 2015-2016, l'INA a déjà une expérience de ces collectes via l'API de la plateforme, ce qui lui permet d'être très réactif sur la collecte de tweets liés aux attentats qui frappent la France. Avec les réseaux socionumériques, les institutions sont face à un volume de données, une volatilité, des formes d'écriture et des circulations entre plateformes qui posent de nouveaux enjeux (auxquels il faut ajouter les enjeux de collecte

via les API ou par le robot Heritrix pour la BnF, le statut public ou privé des comptes et données, etc.). L'interface de l'INA permet l'accès aux millions de tweets préservés lors des attentats de 2015 et 2016, via des recherches par mots-clés, par hashtags, par émojis, etc. (fig. 5). Il y a un saut quantitatif mais aussi qualitatif important pour les chercheurs dans l'approche des données : outre que l'INA propose des recherches par mot-clé, hashtag ou encore par image et émoticône, permettant de ne plus considérer les contenus archivés essentiellement sous l'angle de la recherchabilité textuelle, l'interface est clairement tournée vers la lecture distante, indispensable face à la masse de données en jeu.

Figure 5. Les émojis les plus populaires au sein de la collection Twitter de l'INA portent sur les attentats et ici sur une requête #Jesuischarlie




*Source : © INA.*

Au sein du projet Archives Sauvegarde Attentats Paris (ASAP) lancé en 2016[17] et accompagné régulièrement par les membres du DL Web de l'INA qui organisent formations, sessions de travail communes, etc., se développent ainsi des expériences de lecture distante, témoignant d'un changement d'échelle. Reste que l'expérience menée en 2016 est encore très différente de celle entreprise à partir de 2020 sur la crise de la COVID-19, dans la mesure où les données sont analysées avec les outils proposés par l'institut, induisant des approches et des usages spécifiques. Facilement appropriables, au sein d'une interface *user friendly*, ils permettent des visualisations rapides et efficaces et des premiers pas facilités dans la lecture distante, tout en déléguant à l'interface et donc à l'INA la responsabilité des traitements, ce qui peut maintenir

17. https://asap.hypotheses.org/a-propos.

le chercheur dans une certaine méconnaissance du fonctionnement des outils. En 2020, la situation est différente : les chercheurs reçoivent des données sur la COVID-19, qui ne sont certes pas brutes mais traitables en dehors des interfaces et des sites des institutions (outre qu'elles proviennent de plusieurs institutions d'archivage européennes et permettent donc de croiser ces données). Les projets de la décennie 2010 ont largement préparé ce terrain. Comme le notent Sandras et Stirling (in Beaudouin *et al.*, 2018) au sujet du projet lié à la Grande Guerre développé par Valérie Beaudouin :

> *Ces usages de type « fouille de données » font encore l'objet de débats au niveau national et européen ; des changements se sont produits depuis le début du projet et le contexte juridique reste évolutif. Cette approche par conventionnement a été une manière d'expérimenter de nouveaux modes d'usage par les chercheurs des collections conservées à la BnF, dans un cadre qui respecte le code du patrimoine et la propriété intellectuelle. Le projet a été riche en échanges permettant de mettre en regard la constitution des collections, les outils mis en place pour les exploiter et les besoins de la recherche. Les leçons de ce projet alimenteront les évolutions futures des collections du dépôt légal de l'Internet et leur analyse.*

De cette première période ressortent plusieurs éléments : une évolution des logiques documentaires de collecte comme d'indexation au cours de la décennie ; l'intégration des réseaux socionumériques et la multiplication des collectes spéciales ou d'urgence, au côté des collectes régulières ; des interfaces de recherche plus sophistiquées permises par un travail d'indexation et une réflexion sur la recherchabilité (textuelle, mais aussi par format) ; l'émergence de masses de données qui poussent à la lecture distante ; le souhait des institutions de considérer plus spécifiquement les contenus sous l'angle des données, que ce soit dans le domaine du patrimoine numérisé ou nativement numérique, ce qui engage une évolution radicale du statut du document (Bachimont, 2017) ; l'enrichissement de moyens d'accès, recherchabilité, traitement du web archivé pensés spécifiquement pour les chercheurs ; un besoin d'accompagnement des chercheurs et la volonté des institutions de mieux comprendre ces besoins dans le cadre d'enquêtes ou de séminaires ; des habitudes de travail en commun qui vont se poursuivre et se renforcer, mais qui sont déjà tangibles à la faveur d'événements nationaux et internationaux comme l'accueil de la conférence IIPC par la BnF en 2014 ou l'anniversaire des dix ans du DL Web en 2016. Chercheurs et institutions bénéficient de ces rencontres communes en termes d'échanges de besoins et pratiques, mais tendent aussi à joindre leurs forces autour d'un objet qui peine encore

au mitan des années 2010 à gagner pleinement en légitimité : ces événements communs témoignent auprès des institutions de l'intérêt des chercheurs pour leurs missions et réalisations et rendent visibles les utilisateurs, tout en leur offrant aussi un espace de collaboration qui leur permet de ne pas être isolés et de faire réseau.

Cette tendance à faire réseau se matérialise dans des projets communs, depuis le projet européen RESAW déposé en 2012 et non retenu, mais qui évolue en un réseau aux conférences biennales, jusqu'au projet WARCnet qui réunit plus ou moins les mêmes partenaires et qui est lancé quelques semaines avant le démarrage de la pandémie. Un de ses groupes de travail dédié aux approches transnationales du web archivé par les événements converge assez naturellement vers l'idée d'observer au niveau européen les archives du Web en pleine constitution de la crise de la COVID-19.

## DE LA FOURNITURE DE DONNÉES À LEUR ANALYSE : VERS LA COCONSTRUCTION DEPUIS 2020

Habitués aux restrictions dans le cadre des dépôts légaux et de la conservation par les institutions d'archivage, les membres de l'équipe européenne WARCnet ne pensent pas d'emblée à demander un accès aux données de la COVID-19, mais s'engagent d'abord dans la documentation des collectes, afin de pouvoir les rendre appréhendables pour les chercheurs qui se pencheraient à l'avenir sur ces masses de données.

### Documenter les collectes

La démarche des entretiens oraux réalisés dans le cadre du projet, bien qu'elle puisse sembler s'éloigner du cœur de la réflexion sur les données, est pleinement en prise avec ces enjeux. Elle tire son origine d'une première expérience d'entretiens auprès de l'INA et de la BnF dans le cadre du projet ASAP, précédemment évoqué, afin de comprendre la manière dont les traces numériques des attentats avaient été collectées. Ces témoignages permettent de percevoir la richesse, mais aussi certaines limites des collectes et surtout les choix de curation effectués. Il y a en effet un besoin de contextualiser les archives et l'archivage du Web, pour mener des traitements quantitatifs sur des données dont on ne peut ignorer la constitution, les lacunes ou les biais. Par ailleurs, s'intéresser aux pratiques d'archivage de ces collections permet de rendre visible des infrastructures et des modes de curation qui sont particuliers lorsque l'archivage doit être mené en urgence, puisqu'il sort des logiques des collectes annuelles de préservation large de la « webosphère » nationale.

Les entretiens sur la COVID-19 documentent aussi bien le cas de la Grande-Bretagne que celui de la France, du Luxembourg, des pays nordiques, de l'Irlande ou encore des Pays-Bas. En essayant de définir des questions communes au sein de l'équipe pour mener des entretiens semi-directifs permettant des comparaisons, ce recueil d'entretiens oraux[18] propose une base de documentation au futur chercheur qui souhaiterait analyser les données d'une institution ou encore les croiser avec celles d'autres pays. Il a d'emblée une idée de ce que peut contenir la collection. De plus, se dévoilent au fil de ces entretiens des points communs, et notamment le sentiment des archivistes du Web que documenter cette crise sans précédent peut contribuer à leur mission sociétale et patrimoniale ainsi qu'à la visibilité du rôle des archives du Web. Mais il y a des temporalités à l'œuvre différentes selon les pays : certains commencent les collectes précocement, d'autres les arrêtent au bout de la première vague lorsque les moyens humains ou techniques manquent. On y trouve aussi des explications précises sur les choix des sites archivés, la place accordée ou non aux réseaux socionumériques, les adaptations en cours de la collecte, etc.

Au fil de ces échanges et de cette plongée indirecte dans les contenus, sans pouvoir les voir, puisque beaucoup sont en cours de constitution et ne sont pas encore accessibles aux chercheurs, l'équipe souhaite tester la faisabilité d'accéder à des données.

## Recevoir les données

Les réponses des institutions dépassent les espérances initiales qu'avaient les chercheurs en demandant la fourniture de données de manière assez vague et en ne sachant ni à quelles données ni à quelle masse s'attendre. Ils craignaient un refus pour des raisons législatives, sans compter que cette fourniture demande un travail conséquent de préparation et d'extraction des données. Or, les institutions sont non seulement prêtes à fournir des données dans le cadre d'un projet mené en commun, qui donc fournit déjà un cadre de collaboration, mais aussi parfois à les accompagner de descriptifs pour les rendre plus accessibles aux chercheurs, comme cela a été le cas de la BnF ou de l'INA. Être dans une dynamique européenne permet à chaque institution de pouvoir faire référence à la démarche de leurs homologues à l'étranger et rend la requête moins singulière. Il faut bien sûr passer parfois par des conventions et les différentes institutions sont plus ou moins généreuses en données, exigeantes sur la sécurité ou leur durée de conservation et l'usage qui

18. Publié sous la forme de WARCnet Papers, voir https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports.

peut en être fait. En quelques semaines, nous pouvons toutefois plonger dans toute une série de métadonnées et données dérivées (*seed lists*) qui permettent de prendre la mesure notamment des noms de domaines qui ont été sélectionnés par les institutions d'archivage, mais aussi du démarrage et de l'arrêt (ou pas) de la collecte, du nombre d'URLs conservées et d'autres éléments. Mis en relation avec les résultats des enquêtes orales, ils montrent une forte homogénéité des données (notamment parce que la plupart des institutions passent par des fichiers WARC et le robot Heritrix, à l'exception de l'INA qui dispose de fichiers JSON). Mais il y a aussi une hétérogénéité certaine car toutes les institutions ne documentent pas de la même façon leur collecte et les métadonnées peuvent différer (fig. 6), impliquant un lissage et une mise en interopérabilité par les chercheurs. Présentés à l'occasion de journées mêlant archivistes du Web et chercheurs, les résultats sont aussi importants pour les professionnels qui peuvent y trouver des éléments de comparaison et enrichir leur réflexion sur l'interopérabilité des collections. Il convient de noter que ces lectures distantes appliquées au Web archivé documentent toutefois plus l'archive que le phénomène étudié lui-même.

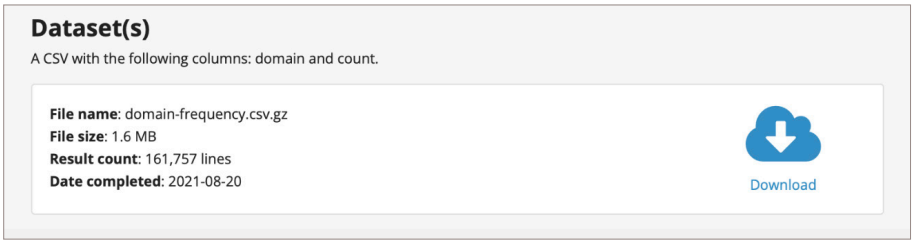
Figure 6. Comparaison des données fournies dans le cadre du projet WARCnet par les différentes institutions (RDL [Royal Danish Library] ; BnF ; NSL [National Széchényi Library, Hungary] ; IIPC ; BnL [Bibliothèque nationale du Luxembourg] ; KB [Koninklijke Bibliotheek, the Netherlands] ; UKWA [UK Web Archive]. Analyse conduite par Karin de Wild et Niels Brügger

| | RDL | BnF | NSL | IIPC | BnL | KB | UKWA |
|---|---|---|---|---|---|---|---|
| Domain Name | | | | | 446 | | |
| TLD | | | | 10708 | 693 | | |
| Exact URL | 17180 | 4598 | 128 | 10734 | 27787 | 603 | 2672 |
| Page title | | | 128 | 10674 | 701 | | 2672 |
| Curator | | | | | | | 2672 |
| Selected on date | 17014 | | | | | | |
| Status | | | | | | | |
| Frequency | | 4598 | | | | | 2672 |
| Theme | | 4598 | | | 700 | | |
| Supplementary archiving | 139 | | | | | | |
| Archiving date | | | | | | | |
| Tweet text | | | | | | | |
| Language | | | | 10693 | | | |
| Twitter ID | | | | | | | |
| Keyword | | 4598 | | | | | |
| Supplementary info | 13375 | 4598 | | 10293 | | | 2672 |
| Country of publication | | | | 10680 | | | |
| URL history | | 533 | | | | | |

*Source : © WARCnet.*

Une seconde étape est franchie ensuite dans le cadre du projet AWAC2 (Aasman *et al.*, 2021a), directement né des premiers résultats européens : en effet, une partie du groupe de travail va saisir l'opportunité d'un appel de l'équipe canadienne d'Archives Unleashed[19]. Celle-ci propose des soutiens à des projets de recherche utilisant une interface spécifique, ARCH, qu'elle a créée (fig. 7) et qui permet d'explorer des données d'Archive-It, liées à Internet Archive. Or la collection internationale qu'a créée l'IIPC sur la COVID-19 a été déposée dans Archive-It et, au terme de discussions avec l'IIPC facilitées par des échanges antérieurs, il est possible d'avoir non seulement accès aux données de la collecte, mais aussi, dans un second temps, au plein texte (ce qui exclut par contre un accès aux archives du Web elles-mêmes et laisse de côté tous les éléments qui ne sont pas textuels, autrement que par une série de données concernant le nombre de fichiers vidéo ou audio, les URLs de ces contenus, etc.).

Figure 7. Interface ARCH développée par l'équipe d'Archives Unleashed et accès via des CSV aux données de l'IIPC (ici le nombre de domaines web conservés)




*Source : © projet Archives Unleashed.*

## Travailler (sur) les données

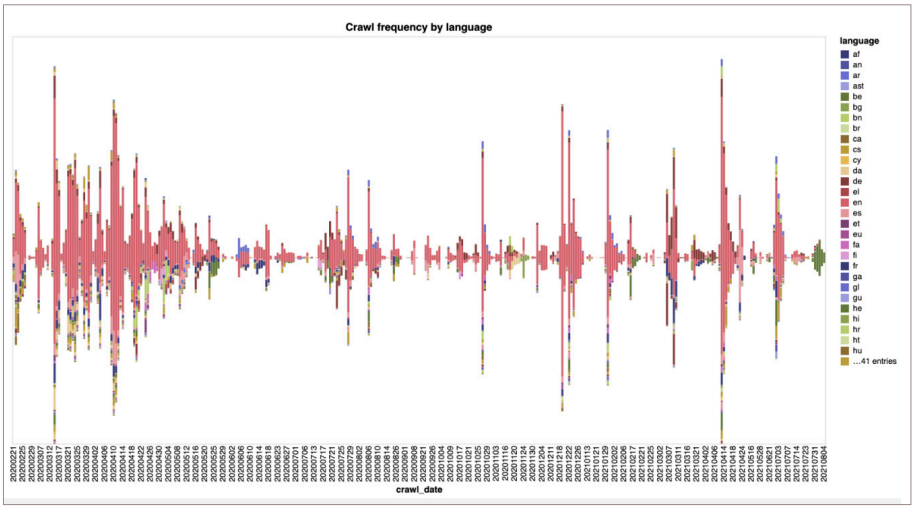
Si la fourniture de données par l'IIPC n'a été possible que par le dialogue engagé depuis plusieurs années, à partir du moment où le projet est lancé, les échanges sont davantage orientés vers l'équipe de recherche Archives Unleashed et ne mettent plus en œuvre un étroit travail commun avec les institutions d'archivage. Toutefois, c'est seulement en raison du travail de préparation en amont, et notamment d'un entretien oral approfondi sur la constitution de la collection, qui permet de comprendre l'origine du corpus (Geeraert et Bingham, 2020), que ce dialogue n'est plus indispensable. Il se poursuit pourtant via des blogposts de l'équipe sur le site de l'IIPC (Aasman

19. https://archivesunleashed.org.

*et al.*, 2021a, 2022) et la participation aux conférences annuelles de l'IIPC pour partager les résultats de la recherche.

Malgré l'interface ARCH (fig. 7) et l'accès facilité aux téléchargements de données qui ont été préparées par les membres d'Archives Unleashed, l'équipe se heurte à plusieurs difficultés : la masse de données à télécharger implique de trouver un hébergement de plus grande capacité que celui des équipements informatiques des membres de l'équipe, mais aussi induit tout un travail de réflexion sur la manière de s'emparer d'une telle collection massive. Nous ne reviendrons pas ici sur les difficultés que peut poser l'appréhension d'un gros corpus multilingue (fig. 8) ou sur le bruit que l'on trouve dans les résultats, il nous semble plus intéressant d'insister sur les dispositifs d'approches de ces données : ils passent essentiellement par des séances collectives de travail et des datathons internes (Aasman *et al.*, 2021b), qui s'appuient sur l'expérience menée par l'équipe d'Archives Unleashed, tout en l'adaptant (Fritz, Milligan, Ruest, Lin, 2021). Il s'agit, par exemple, de définir des *workflows* de traitement des données ou encore de choisir les meilleurs outils et algorithmes pour la fouille de données (Aasman *et al.*, 2022), question qui ne se pose pas quand les corpus sont mis à disposition au sein des institutions, puisque les outils sont alors contraints.

Figure 8. Fréquence des captures par langue au moyen de Pandas et de bibliothèques Altair pour Python. Réalisation par Frédéric Clavert




*Source : © Frédéric Clavert.*

Il convient également de souligner les nombreuses étapes techniques traversées par ces données, depuis leur sélection par les institutions nationales, leur collecte par l'IIPC (les institutions remontent des URLs mais pas les contenus eux-mêmes), leur mise à disposition dans Archive-It puis à travers une interface spécifique ARCH (qui fait aussi des prétraitements sur les fichiers WARC pour proposer des entrées par nom de domaine, types de fichiers MIME, etc.), enfin, les téléchargements du corpus et sa division en sous-corpus. Une telle chaîne de transformations, de compétences et d'outils pose la question de la possibilité de capitaliser sur les méthodologies et de sa reproductibilité, mais aussi celle des biais introduits au cours des traitements.

## Coconstruire les données

Si l'exemple de WARCnet permet de voir une coconstruction à l'œuvre, à la fois en termes de documentation des collectes et de fournitures de données, engageant aussi à de nouvelles pratiques collaboratives et interdisciplinaires, des temps d'échanges sont aussi pensés par les institutions et notamment au sein d'expériences de datathons, comme ceux qu'organise Marta Severo avec l'INA. De même, en 2016 à la BnF :

> *Un hackathon de 24 heures est organisé : la BnF ouvre ses portes à des équipes de développeurs, chercheurs, étudiants, designers et autres passionnés pour imaginer et prototyper de nouveaux projets autour de ses collections numériques. Cette initiative, reconduite en 2017 et 2018, fait apparaître l'intérêt d'un accompagnement de proximité autour de l'usage des données de Gallica, de Data.bnf.fr ou autres, afin de donner aux usagers toutes les clefs dont ils ont besoin pour les réutiliser à pleine puissance. En 2017, lors de la deuxième édition du hackathon, un site dédié à la documentation de ces accès techniques est créé : BnF API et jeux de données. (Bermès, 2020, p. 58-59).*

À la faveur de ces hackathons et datathons organisés par les institutions, outre que se créent des rapports de collaboration entre professionnels de l'archivage et de la conservation et équipes de recherche, les participants peuvent davantage explorer les fonds, entrer dans leur constitution et leurs coulisses, exprimer des besoins et découvrir de nouvelles pistes d'investigation, tout en bénéficiant de la mise en commun et de la découverte de projets parallèles.

Une nouvelle expérience d'accompagnement va couvrir, elle, toute la chaîne de traitement ou presque des données, de la constitution du corpus à son analyse et à sa dissémination, dans le cadre du DataLab de la BnF en 2021-2022 (projet BUZZ-F). Alors que pour la COVID-19 il y a moins

d'échanges interprofessionnels, une fois les données fournies, le corpus ayant été constitué par les archivistes et servant de base de départ à la recherche, il est intéressant de voir ici un exemple de coconstruction du corpus mais aussi de l'analyse. Cette collaboration d'un an bénéficie de l'appui de l'équipe du DL Web ainsi que d'un ingénieur d'Huma-Num, associé à l'accompagnement des chercheurs au sein du BnF DataLab. La recherche souhaite analyser des phénomènes de viralité (mèmes, popularité de certains hashtags, contenus qui « font le buzz » en ligne) à travers les archives du Web[20]. Or il faut composer avec une collection d'archives à la BnF inégalement indexée plein texte[21]. L'idée est de commencer par explorer les traces du Lip Dub et du Harlem Shake qui remontent à la première moitié des années 2010, pour croiser ces problématiques de recherchabilité entre collections indexées et non indexées. Les discussions constantes et les va-et-vient permettent d'imaginer ensemble des stratégies d'identification des contenus et d'extraction de mots-clés via les URLs. L'apport pour l'équipe de recherche est certain, puisqu'elle n'aurait pas pu constituer un corpus un tant soit peu complet sur la thématique et identifier des contenus non indexés sans le travail des équipes du DL Web, ce que la recherche par URLs en interne va permettre, ou encore prendre la mesure des doublons, comparer les contenus archivés et ceux toujours présents en ligne. Mais il y a aussi un apport pour la BnF en termes d'ingénierie de projet, de soutien technique et documentaire (Pailler et Faye, 2022), sans compter la réalisation de visualisations par l'équipe BnF / Huma-Num (fig. 9).

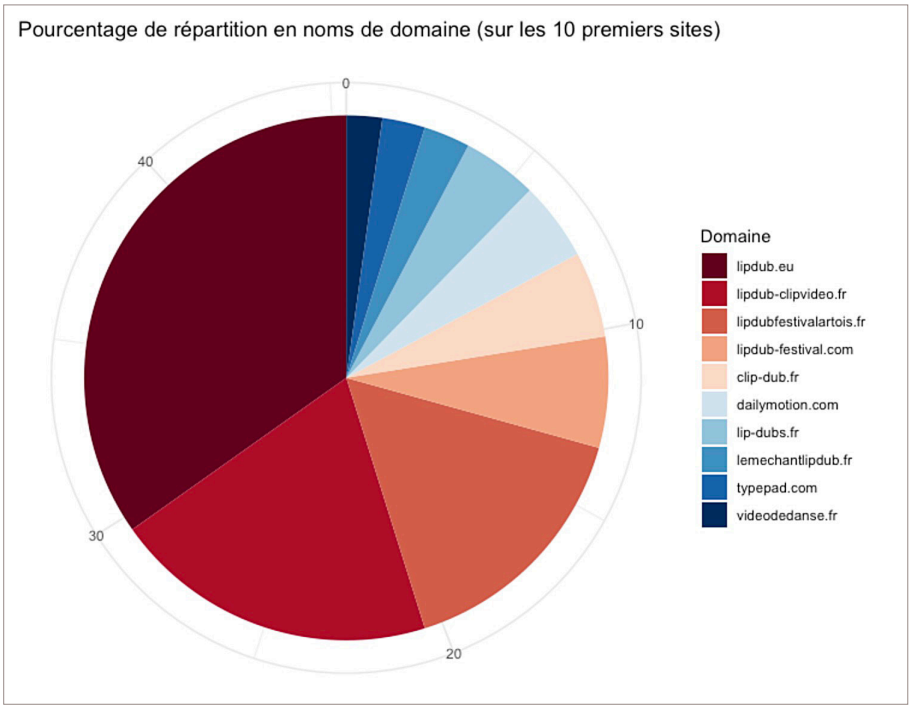
Aussi depuis 2020, à l'occasion des recherches transnationales menées sur les données de la pandémie, à partir de corpus constitués, mais aussi d'un projet de recherche plus national de création de corpus mêlant lecture proche et distante, des tendances en préfiguration lors de la décennie précédente se confirment : une entrée par les données du Web archivé, que ce soit des métadonnées ou des données dérivées et la mise à disposition de celles-ci en dehors de l'institution ; la diversification des outils de traitement (également intégrés au sein des laboratoires, pour permettre davantage de fouilles de données, de modélisation thématique, d'études de réseaux ou encore visuelles, etc.) ; une documentation renforcée des collectes par les archivistes mais aussi par les chercheurs ; des analyses menées en commun, qui se traduisent aussi par des publications et présentations communes, alors que la

20. En prenant pour appui les besoins d'un projet d'étude transnational en cours HIVI (A history of online virality), soutenu par le Fonds national de la recherche (FNR) au Luxembourg (C20/SC/14758148). En ligne : https://hivi.uni.lu/.

21. Indexation pour les années 1990, la période des attentats et pour les actualités en ligne, et partiellement pour la crise COVID. Le reste des collections n'est pas encore indexé plein texte, la BnF ayant commencé une indexation de ses fonds en entrant par des collectes thématiques.

décennie précédente était déjà à la collaboration, mais moins à la coconstruction des recherches. Cette évolution est facilitée par les étapes précédentes qui ont permis de construire un langage et des enjeux communs, de mieux comprendre les attentes, pratiques, temporalités, et contraintes respectives.

Figure 9. Recherche dans les URLs des archives du Web de la BnF du terme lipdub par Alexandre Faye (BnF) et Antoine de Sacy (Huma-Num)




*Source : © BnF.*

## CONCLUSION : VERS ENCORE PLUS DE PARTAGE ?

La fourniture de données invite à de nouvelles pratiques et de nouveaux services et collaborations entre professionnels de l'archivage du Web et chercheurs, qui passent par l'accompagnement technique et l'outillage des recherches (voir le BnF DataLab et sa collaboration étroite avec les chercheurs ; l'accès à la collection COVID-19 de l'IIPC par l'interface ARCH développée dans le monde de la recherche, etc.) mais aussi par un accompagnement cognitif (par exemple au travers de la vaste campagne d'entretiens

oraux menée au sein du projet WARCnet auprès des archivistes du Web, voir Geeraert, Winters *et al.*, 2021).

Les « lieux de savoir » en se faisant toujours davantage lieux de fourniture de données et de services contribuent au développement de compétences en humanités numériques en interne comme à destination des chercheurs et jouent un rôle renforcé dans les recherches. Ils engagent à des partenariats entre professionnels des archives et monde académique, tout en posant aussi la question de l'évolution des métiers et compétences, des moyens donnés aux bibliothèques pour un tel accompagnement, ainsi que des contributions respectives et de la reconnaissance de cette coconstruction.

Cette évolution renforce la place des institutions patrimoniales comme des lieux d'ingénierie, de médiation, d'accompagnement de la recherche. La question peut même être poussée plus loin : ils se font toujours plus lieux de recherche. Les réalisations, par exemple dans Buzz-F, interrogent la césure à effectuer entre ingénierie et recherche car ces outils institutionnels et ces expertises en humanités numériques n'ont rien à envier aux outils développés dans le cadre de laboratoires académiques. Les interventions pour la réalisation de corpus des archivistes du Web ont évidemment une influence sur la manière dont les chercheurs vont mener leurs recherches. La question de la coauctorialité des recherches, de la reconnaissance des contributions respectives, est donc un enjeu.

L'évolution vers les données pose aussi la question de la capacité des institutions d'archivage, au fur et à mesure que les demandes vont se multiplier, à pouvoir répondre à ces besoins. Il y a ici des enjeux évidemment humains, mais aussi la question des dispositifs permettant de reconnaître (concrètement et symboliquement) ce travail qui n'est pas invisible certes, mais chronophage et hautement technique, d'accompagnement des chercheurs. Le coût de la fourniture des données est une question qui pourrait sans doute se poser dans les années à venir. En engageant de nouveaux dispositifs de collaboration avec les chercheurs, les archivistes du Web ont évidemment des avantages et notamment celui de saisir plus précisément les besoins à mettre en œuvre pour les chercheurs, mais se pose ici aussi pour eux la question du développement d'outils spécifiques, sur mesure, ou au contraire la recherche « d'outils et services plus universels » répondant à une demande large.

Avec le développement du FAIR data[22] (Mons, 2018), appuyé notamment sur l'accès ouvert et qui pose explicitement la question de la recherchabilité, de l'accès, de l'interopérabilité, et de la réutilisation des données, il reste aussi

22. L'acronyme FAIR repose sur le fait que les données soient *findable*, *accessible*, *interoperable* et *reusable*.

à se pencher sur la capacité à pouvoir partager les corpus, capitaliser sur les résultats et méthodologies existants, et les mettre à disposition dans le cadre de travaux ultérieurs.

Des besoins de formation sont également tangibles et il convient de se demander où mener ces expériences et transferts de littératie numérique, qui doivent prendre appui sur les données et devraient donc être certainement menés au plus près des données. Ce n'est pas encore toujours possible en raison des accès limités aux archives du Web dans les universités[23], mais un projet comme ResPaDon[24], qui vient de se clôturer, a œuvré en ce sens côté BnF avec l'université de Lille et plus largement avec la communauté académique, tandis que l'INA a déjà développé des accès dans quelques bibliothèques universitaires. On ne peut que souhaiter la généralisation des accès aux archives du Web dans les établissements de l'enseignement supérieur et de la recherche, tout en notant que ce Web archivé des données dépasse aussi la question de la seule recherche académique pour pouvoir toucher d'autres publics.

## BIBLIOGRAPHIE

Aasman, S., Brügger, N., de Wild, K., Clavert, F., Gebeil, S., Schafer, V. (2021a). Analysing Web Archives of the COVID-19 Crisis through the IIPC collaborative collection: early findings and further research question. IIPC netpreserve.org. Repéré à : https://netpreserveblog.wordpress.com/2021/11/02/analysing-web-archives-of-the-covid-19-crisis-through-the-iipc-collaborative-collection-early-findings-and-further-research-questions/.

Aasman, S., Bingham, N., Brügger, N., de Wild, K., Gebeil, S., Schafer, V. (2021b). Chicken and Egg: Reporting from a Datathon Exploring Datasets of the COVID-19 Special Collections. *WARCnet Paper*. Repéré à : https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Aasman_et_al_Chicken_and_Egg.pdf

Aasman, S., Brügger, N., de Wild, K., Clavert, F., Gebeil, S., Schafer, V., Sirajzade, J. (2022). Studying Women and the COVID-19 Crisis through the IIPC Coronavirus Collection, IIPC. Repéré à : netpreserve.org, https://netpreserveblog.wordpress.com/2022/12/20/studying-women-and-the-covid-19-crisis-through-the-iipc-coronavirus-collection/.

Bachimont, B. (2017). *Patrimoine et numérique. Technique et politique de la mémoire*. Bry-sur-Marne : INA.

23. Il y a toutefois des lieux de consultation en région. Voir par exemple pour l'INA : https://www.ina.fr/institut-national-audiovisuel/en-regions.
24. https://respadon.hypotheses.org.

Beaudouin, V., Chevallier, P., Maurel, L. (éd.) (2018). *Le web français de la Grande Guerre. Réseaux amateurs et institutionnels*. Nanterre : Presses universitaires de Nanterre.

Beaudouin, V. (2019). Comment s'élabore la mémoire collective sur le web ? Une analyse qualitative et quantitative des pratiques d'écriture en ligne de la mémoire de la Grande Guerre. *Réseaux*, volumes II-3 (n° 214-215), p. 141-169. DOI : https://doi.org/10.3917/res.214.0141.

Bermès, E. (2020). *Le numérique en bibliothèque : naissance d'un patrimoine : l'exemple de la Bibliothèque nationale de France (1997-2019)*. Paris : École nationale des chartes. Repéré à : https://theses.hal.science/tel-02475991/document.

Bingham, N. (2020). IIPC content development Group's activities 2019-2020. Repéré à : https://netpreserveblog.wordpress.com/2020/07/01/iipc-content-development-groups-activities-2019-2020/.

Brügger, N. (2020). Welcome to WARCnet. *WARCnet Paper*. Repéré à : https ://cc.au.dk/fileadmin/user_upload/WARCnet/1.Bru__gger_Welcome_to_WARCnet.pdf.

Brügger, N. (2018). *The archived Web. Doing history in the digital age*. Cambridge, MA: MIT Press.

Carlin, M. & Laborderie, A. (2021). Le BnF DataLab, un service aux chercheurs en humanités numériques. *Humanités numériques,* n° 4. DOI : https://doi.org/10.4000/revuehn.2684.

Clavert, F. (2018). Face au passé : la Grande Guerre sur Twitter. *Le Temps des médias*, n° 31, p. 173-186. DOI : 10.3917/tdm.031.0173.

Fritz, S., Milligan, I., Ruest, N., Lin, J. (2021). Fostering Community Engagement through Datathon Events: The Archives Unleashed Experience. *Digital Humanities Quaterly*, 15, 1. Repéré à : http://www.digitalhumanities.org/dhq/vol/15/1/000536/000536.html.

Geeraert, F., Winters, J. *et al.* (18 juin 2021). Representation, participation and inclusivity: European web archives collecting the digital traces of COVID-19. Communication présentée lors de la IVe conférence RESAW, Université du Luxembourg. Repéré à : https://www.resaw2021.net/programme/.

Geeraert, F. et Bingham N. (2020). Exploring special web archives collections related to COVID-19: The case of the IIPC collaborative collection. WARCnet Paper. Repéré à : https://cc.au.dk/fileadmin/user_upload/WARCnet/Geeraert_et_al_COVID-19_IIPC__1_.pdf.

Merzeau, L., Mussou, C. (2017). L'expérience des ateliers du dépôt légal du web de l'INA. Repéré à : https://webcorpora.hypotheses.org/302.

Mons, B. (2018). *Data Stewardship for Open Science. Implementing FAIR Principles*. Boca Rota: CRC Press.

Moiraghi, E. (2018). Le projet Corpus et ses publics potentiels. Paris, Bibliothèque nationale de France. Repéré à : https://hal-bnf.archives-ouvertes.fr/hal-01739730.

Moretti, F. (2007). *Graphs, Maps, Trees: Abstract Models for Literary History*. Londres et New York, Verso. Repéré à : https://docdrop.org/static/drop-pdf/Pages-from-Moretti-2007---Graphs-Maps-bkBff.pdf.

Pailler, F., Faye, A. (9 décembre 2022). Présentation du projet Buzz-F. Présentation des travaux des projets lauréats de l'Appel à projet BnF DataLab/Huma-Num, Paris.

Pomerantz, J. (2015). *Metada.* Cambridge, MA: The MIT Press.

Roustan, M., Monjaret, A. (dir.) (2016). *La recherche dans les institutions patrimoniales : sources matérielles et ressources numériques*. Villeurbanne : Presses de l'Enssib.

Ruest, N., Fritz, S., Deschamps, R., Lin, J. & Milligan, I. (2021). From archive to analysis: accessing web archives at scale through a cloud-based interface. *International Journal of Digital Humanities*. DOI : https://doi.org/10.1007/s42803-020-00029-6.

Sandras, A., Stirling, P. (2018). Constituer une archive du Web de la Grande Guerre et la rendre accessible aux chercheurs. In Beaudouin, V., Chevallier, P., Maurel, L. (dir.). *Le web français de la Grande Guerre. Réseaux amateurs et institutionnels*. Nanterre : Presses universitaires de Nanterre, p. 91-105. Repéré à : https://books.openedition.org/pupo/22422.

Schafer, V., Musiani, F., Borelli, M. (2016). Negotiating the Web of the Past. *French Journal for Media Research*, 6. Repéré à : http://frenchjournalformediaresearch.com/lodel/index.php?id=963. hal-01654218.

Treleani, M. (2017). *Qu'est-ce que le patrimoine numérique ? Une sémiologie de la circulation des archives*. Lormont : Le Bord de l'eau.